U.S. Opioid Prescriber Analysis

**Introduction:**

The opioid problem in the United States has been ongoing since the 1990s and has only accelerated in recent years. Opioids are “a class of drugs naturally found in the opium poppy plant and that work in the brain to produce various effects, including the relief of pain” (Johns Hopkins). Even with ample public information about the severe threat opioids pose, overdose deaths due to opioids have quadrupled since 1999 (CDC). Data mining and analysis might be helpful to find insights into why overdose deaths are occurring and possible ways to help prevent them, whether that be from the source, the doctors prescribing them, or the vulnerable communities that are receiving these opioids.

**Purpose:**

Our goal is to classify and evaluate opioid prescribers to highlight possible anomalies, identify high risk areas and prescriber types to assist early intervention programs. One of our data mining problems is to discover which doctors prescribe the most opioids. The information will help us understand some aspects of the epidemic. We will also explore which opioids are most likely prescribed by a particular type of prescriber. Additionally, we hope to identify prescribers/doctors who appear to be prescribing opioids at a higher rate than other doctors of their type. To find these insights, we will be using Association Rule Mining, Clustering and Classification Algorithms to perform our analysis.

**Data Description:**

For our project, we will analyze United States Opiate Prescriptions. Kaggle hosts a subset of the dataset provided by the Centers for Medicare & Medicaid Services (CMS), a part of the U.S. Department of Health and Human Services. The original dataset is accessed on the cms.gov website and has over 23 million records for each year starting from 2013. The Kaggle dataset “contains summaries of prescription records for 250 common opioid and non-opioid drugs written by 25,000 unique licensed medical professionals in 2014 in the United States for citizens covered under Class D Medicare and some metadata about the doctors themselves.” To make our analysis timelier and more relevant, we pulled data for 2019 from the cms.gov website.

**Data Processing:**

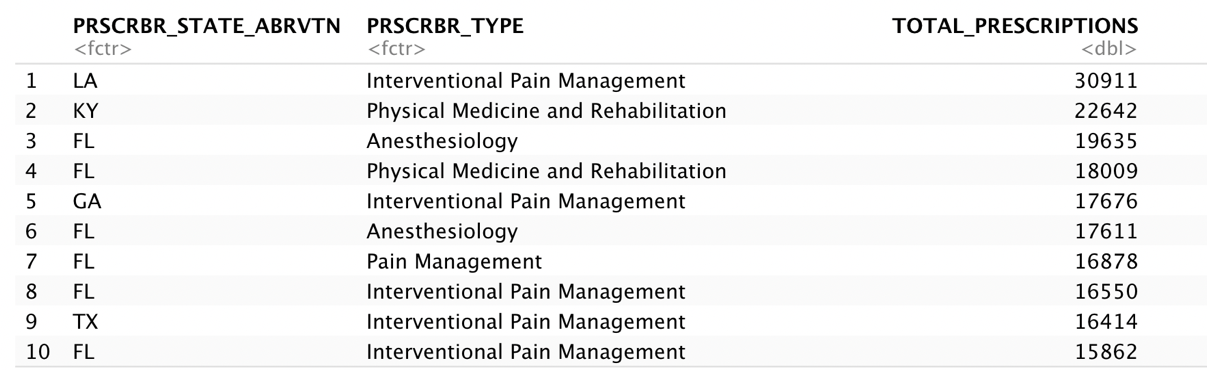
As we begin our analysis we first need to process and clean our data. We leveraged the data consolidation script found on Kaggle with the original subset of data, on the new data we pulled for 2019. The final data set consists of 971,968 observations and 254 variables, in it are essential variables such as Prescriber NPI Number, State, Specialty (dentist, internal medicine, optometry, etc.), and fields for the 250 common opioid and non-opioid drugs.

As we began comparing the different drug types, we found that we needed to use the generic drug name, not the brand name, to analyze the data most accurately. In total, we counted 10 total generic drug names for opioids. We created a Boolean label Opioid Prescriber, which was assigned a value of TRUE if the prescriber prescribed opioid drugs at least once in 2019. Before we started building different models, we converted the Opioid Prescriber variable to factor.

**Descriptive Analytics:**

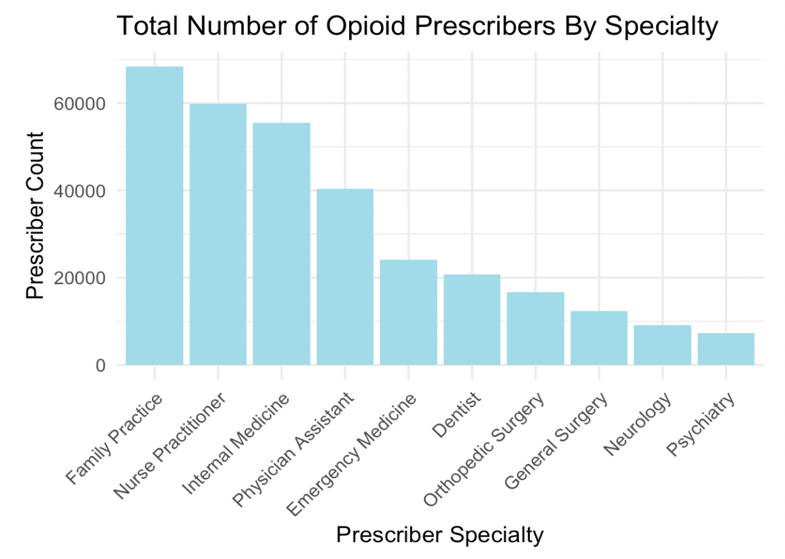
The data set includes prescriber records from 183 different specialties. Approximately 59% of the prescribers did not prescribe any opiate drugs in 2019.

We first reviewed the total number of opioid prescriptions wrote by each prescriber and found out the distribution of that data is extremely right skewed, with a median of 49.0 and mean of 152.6. Among the opioid prescribers, half of them prescribed opiate drugs less than 50 times a year. Below is the information for the top 10 opioid prescribers ranked by total number of opioid prescriptions.

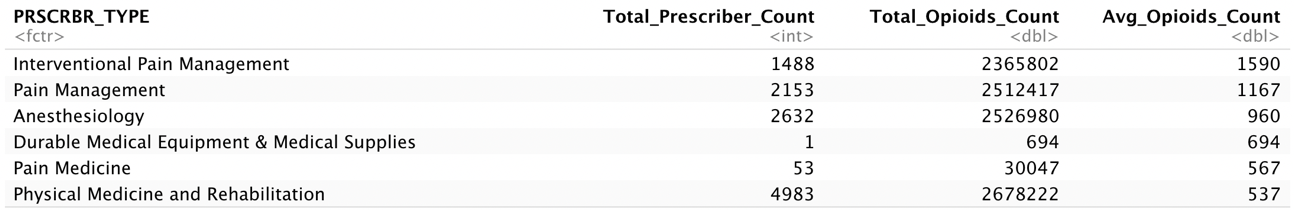


As we can see from the table above, one individual wrote more than 30,000 opioid prescriptions in 2019, which was significantly higher than all other prescribers in the country. Most people on the list were practicing interventional pain management or regular pain management, and they may need to give large amounts of pain medicines to patients every day. In addition, we found that these top 10 prescribers were all from states in the Southern United States, which only has two of the top five U.S. states based on population size.

We then summarized the total number of opioid prescribers and opioid prescriptions by specialty. The plot below displays the top 10 specialties with the highest total number of opioid prescribers.



The table below lists the top 6 specialties with the highest average opioid prescriptions written by an individual prescriber.



Not surprisingly, the top two prescriber specialties by average opioid prescription count are interventional pain management and regular pain management. We also found another unusual record on the list above. Durable medical equipment & medical supplies specialty had only one opioid prescriber in 2019, but the average number of opioid prescriptions this person wrote ranked the fourth among all specialties.

In addition, we reviewed the drug data and found among the 10 opiate drugs in our data set, Hydrocodone Acetaminophen, Tramadol, Oxycodone Hcl Acetaminophen and Oxycodone were most prescribed in 2019. Each opiate drug variable includes numeric values and is extremely right skewed with mean well above the median.

**Association Rule Mining:**

We first performed Association Rule Mining to explore relations between prescriber type and drug information. We wanted to find out which opioids are most likely to be prescribed by certain types of medical professionals. To achieve the goal, we created a subset of data that only included information for opioid prescribers and opiate drugs. Since each opiate drug variable includes numeric values and is right skewed, we first converted them to factors by discretizing those variables into customized bins. We tried more than four different ways to perform discretization, and the final bin labels we chose were "0-10","10-500","500-1000","1000-2000","2000-4000" and "Above 4000".

Next, we used the arules package and the apriori function in R to generate rules that have Prescriber Type on the left-hand side. We initially set the support threshold at 0.10, confidence threshold at 0.80 and set rule length equals to 2, but not many interesting rules were generated with those criteria. Therefore, we reduced the support to 0.03 and changed the confidence threshold to 0.70, and below are the five most interesting rules we found. The rules were generated within a second and filtered using the lift value.

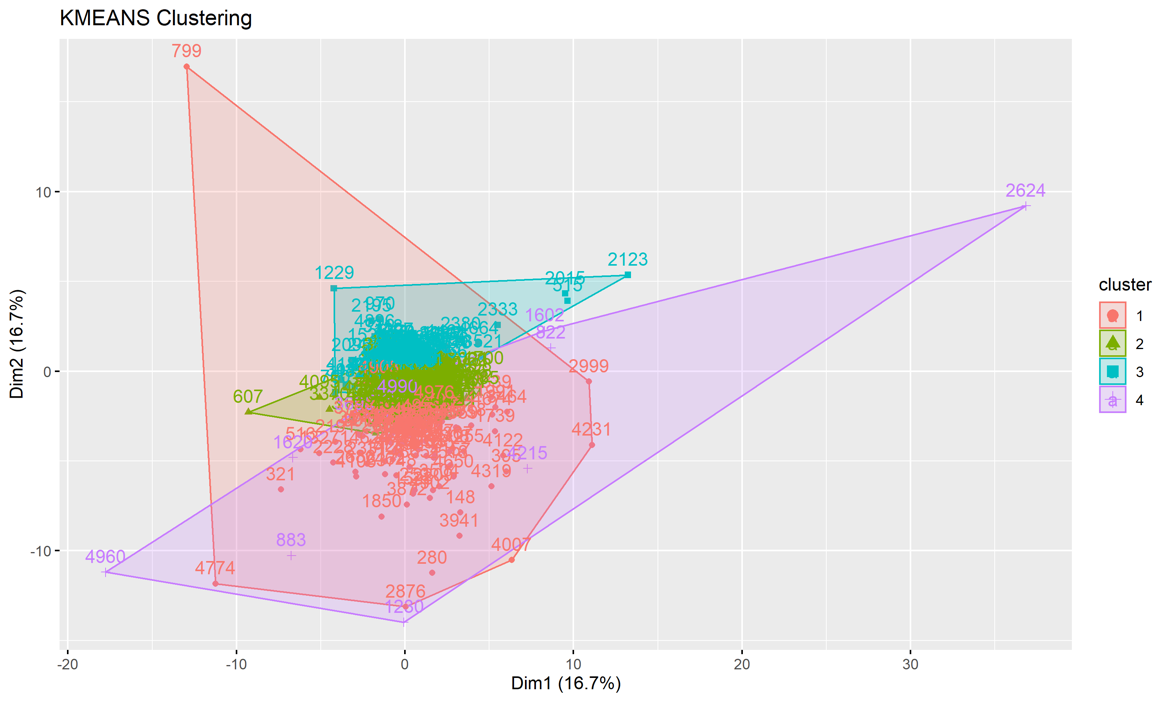
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lhs (Prescriber Type)** |  | **Rhs (Drug Information)** | **supp** | **conf** | **lift** |
| Family Practice | => | TRAMADOL.HCL=10-500 | 0.1371 | 0.8050 | 1.4827 |
| Family Practice | => | HYDROCODONE.ACETAMINOPHEN=10-500 | 0.1263 | 0.7413 | 1.2207 |
| Internal Medicine | => | TRAMADOL.HCL=10-500 | 0.1080 | 0.7824 | 1.4409 |
| Orthopedic Surgery | => | HYDROCODONE.ACETAMINOPHEN=10-500 | 0.0306 | 0.7393 | 1.2174 |
| Emergency Medicine | => | HYDROCODONE.ACETAMINOPHEN=10-500 | 0.0434 | 0.7212 | 1.1876 |

Each rule above has a lift value greater than 1. These rules suggest that if the prescriber is practicing family practice, he or she is around 81% likely to prescribe Tramadol 10-500 times a year, and around 74% likely to prescribe Hydrocodone Acetaminophen 10-500 times a year. If the prescriber is practicing internal medicine, he or she is around 78% likely to prescribe Tramadol 10-500 times a year. If the prescriber is practicing orthopedic surgery or emergency medicine, he or she is around 73% likely to prescribe Hydrocodone Acetaminophen 10-500 times a year.

The results from the Association Rule Mining may be used by the government authority to monitor which specialties are most likely to prescribe opioids at high dosages. We also reviewed rules that have the State name on the left-hand side but did not find any interesting results. Association Rule Mining may not be a good technique to analyze relations between State and drug name/prescription amount.

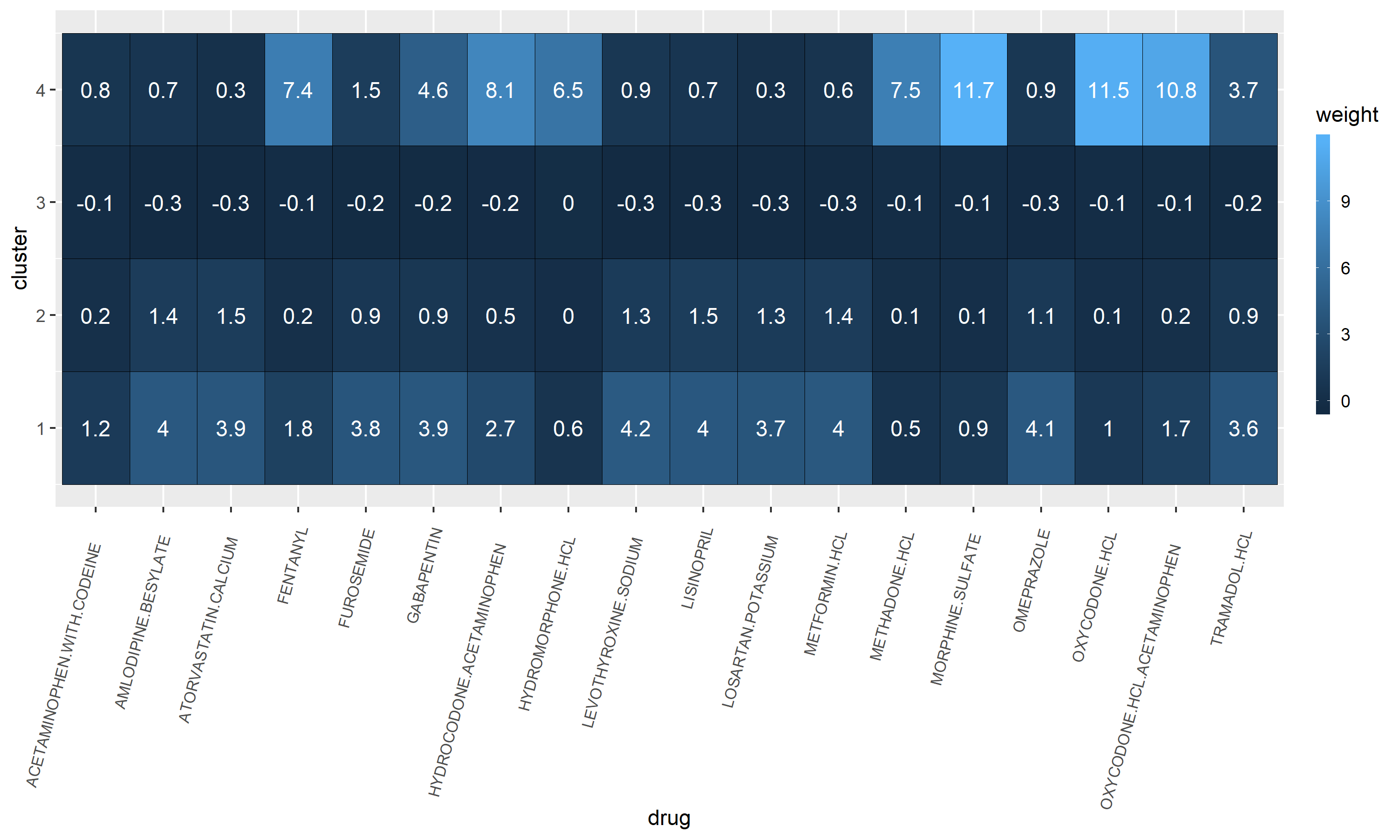
**Clustering Analysis:**

For the final model we did a KMeans Clustering model to see how each of the prescriber’s prescribing behaviors make them similar or different than other prescribers. KMeans is able to do this by finding patterns within the data to automatically cluster observations together to minimize difference between all the observations within each cluster, while maximizing the differences between each cluster. To simplify the data for the model, a subset of the data was used to limit analysis to the top 20 prescriber types which kept around 90% of the data. Additionally, the top 9 commonly prescribed opioids and non-opioids among the 255 drugs were used as the features for the model.



To determine what would be the optimal number of clusters to use for the model, a Silhouette plot was created, above. The model determined 4 clusters to be the optimal number based on the data. The predicted cluster data is added to the original data, and used to create a heatmap to help interpret how each of the clusters were created. Based on heatmap the cluster appears to be created based on the following criteria:

* **Cluster 1** – Prescribers that prescribe an average number of opioids, and a high number of non-opioid drugs
* **Cluster 2** – Prescribers here prescribe a low number of opioid drugs, and an average amount of non-opioid drugs
* **Cluster 3** – Prescribers that prescribe a very low amount of opioid and non-opioid drugs
* **Cluster 4** – This group of prescribers prescribe a very high number of opioid drugs and a very low amount of non-opioid drugs



**Classification Analysis:**

**Data Prep:**

The algorithms selected for the Classification Analysis were C5.0 Decision Tree and Naïve Bayes Classification in R. Data preparation was needed before training the algorithm. The target variable, Opioid Prescriber, was converted to a factor. Also, observations without Prescriber Type and State were removed from the analysis. A validation set was created using 20% of the training data and acquired through random sampling. Lastly, there were some challenges creating a balanced validation set and training set when using the Prescriber Type and Prescriber State attributes. The model failed at predicting on the validation set if it had Prescriber Types and States that were not part of the training set. Therefore, the validation set was filtered so that it consists of only Prescriber Types and States included in the training set.

**Decision Tree Analysis:**

The C5.0 algorithm was able to correctly classify 100% of the training data in about 16 minutes. The average tree size was 10.2. The splitting attributes are Fentanyl, Hydrocodone Acetaminophen, Hydromorphone HCL, Methadone HCL, Tramadol HCL, Topiramate, Oxycodone HCL, Oxycodone HCL Acetaminophen, Acetaminophen with Codeine, and Morphine Sulfate. 8/9 drugs are opioids while one, Topiramate, is not an opioid. The eight opioids are the most important deciding attributes for opioid prescribers. The model performed exceptionally well on the validation set also with an accuracy of 100%.

Table

Description automatically generatedText

Description automatically generated

Text

Description automatically generated Table

Description automatically generated

**Naïve Bayes Analysis:**

The Naive Bayes model produced the conditional probabilities associated with attributes and the target variable, Opioid Prescriber. The Naive Bayes model utilized the Prescriber Type, Prescriber State, and drug attributes to train the model. The a priori probabilities for Opioid Prescriber are as follows: False = 0.587, True = 0.413. When reviewing the conditional probabilities by State, California had the highest conditional probability for Opioid Prescriber = True and Opioid Prescriber = False. California is a large state; this makes sense conceptually. American Samoa had the lowest conditional probability for Opioid Prescriber = True and Opioid Prescriber = False. States with relatively high conditional probabilities for Opioid Prescriber = True are Texas, Florida, New York, Pennsylvania, Ohio, Illinois, Michigan, and North Carolina.

Reviewing conditioning probabilities by Prescriber Type reveals Family Practice had the highest conditional probability for Opioid Prescriber = True. Internal Medicine, Physician Assistant, Orthopedic Surgery, and General Surgery also revealed relatively high conditional probabilities for Opioid Prescriber = True. Dentists had the highest conditional probability for Opioid Prescriber = False and, interestingly, a relatively high conditional probability for Opioid Prescriber = True.

The Naive Bayes model did not perform as well as the Decision Tree model on the validation set and produced an accuracy of 72.8%. However, Naive Bayes ran much faster than the Decision Tree.

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Table

Description automatically generated with low confidence

**Support Vector Machine:**

In addition to Decision Tree and Naïve Bayes, we also attempted a support vector machine model. Our goal with this model was to correctly classify opioid prescribers, much like we did with the other two classification models. We used the same training and validation sets for our SVM, however unlike the other models we were unable to get the SVM to complete. Initially, when building the model with our entire testing dataset, it ran for over 24 hours but never completed. Once we realized that it would not be completed in a reasonable amount of time, we reduced the training data. To do that we sampled 20% of the existing training data and tried to re-train a new SVM with that new, smaller subset. Unfortunately, that model also ran for over 24 hours, but did not complete. Since both attempts with our support vector machines did not complete, we were unable to evaluate and compare it with our other models, and ultimately unable to include it in our analysis.

**Limitations:**

There were a few factors in our analysis which we did not have the time to compensate for. One was the population of each state. It is very difficult to compare state overdose and opioid prescriptions state-by-state without normalizing by the state populations. It is much different for there to be 10,000 overdose deaths in California, with a population of almost 40 million people, than in Vermont, where there are only slightly more than 600,000 residents (World Population Review). Due to time constraints, we were not able to normalize our data against state population size.

During our analysis, we also realized there are several larger socio-economic issues that play a significant part in opioid prescriptions and overdoses. Levels of poverty, race, ethnicity, work force participation, levels of education, and access to healthcare are just a few of the factors that affect opioid overdoses. (Heyman et al., 2019).

**Conclusion:**

The table below summarizes the performance of our classification models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Parameters** | **Accuracy** | **Precision** | **Recall** | **Time to Train** |
| Classification – Decision Tree | Trials = 6, Global Pruning, minCases = 30, CF = 0.75 | 100% | 1.0 | 1.0 | 16 minutes |
| Classification - Naïve Bayes | Laplace = 1, usekernel = TRUE, usepoisson = TRUE | 72.8% | 0.70 | 0.93 | Less than one minute |

In conclusion, all models provided crucial and consistent information on the Opioid Prescriber dataset. We found relations between prescriber type and opioid drug through Association Rule Mining, which can be used to monitor which specialty prescribe certain drugs at specific dosages. Naïve Bayes highlighted high-risk States as well as Prescriber Types. Association Rule Mining and the Decision Tree provided more information on the commonly prescribed opioids. Family Practices, Internal Medicine prescribers, Physician Assistant, Orthopedic Surgeon clinics, and General surgery clinics should be monitored due to high amounts of Opioid prescriptions. Furthermore, access to rehabilitation should be made readily available specifically in high-risk States like Florida, Texas, California, and North Carolina.

Sources

Heyman, G. M., McVicar, N., & Brownell, H. (2019). Evidence that social-economic factors play an important role in drug overdose deaths. *The International journal on drug policy*, *74*, 274–284.

https://doi.org/10.1016/j.drugpo.2019.07.026

*What Are Opioids?* (n.d.). Hopkinsmedicine.Org. https://www.hopkinsmedicine.org/opioids/what-are-opioids.html

*Understanding the Epidemic | Drug Overdose | CDC Injury Center*. (n.d.). CDC.Gov. https://www.cdc.gov/drugoverdose/epidemic/index.html#:%7E:text=The%20first%20wave%20began%20with,overdose%20deaths%20involving%20heroin4.

World Population Review. (n.d.). *US States - Ranked by Population 2022*. https://worldpopulationreview.com/states